British Journal of Clinical Psychology (1997), 3 6, 609-617 *Printed in Great Britain*

Assessing the reliability and abnormality of subtest differences on the Test of Everyday Attention

J. R. Crawford" and J. Sommerville
Department of Psychology, Kings College, University of Aberdeen, AB9 2UB, Scotland

Ian H. Robertson
MRC Applied Psychology Unit, Rehabilitation Research Group, Addenbrookes Hospital, Cambridge, CB2 2QQ UK

Objectives. To assist clinicians with the analysis of an individual's profile of subtest strength and weaknesses on the Test of Everyday Attention (TEA).

Design. The study applied psychometric methods for the quantitive analysis of subtest profiles (Silverstein, 1982, 1984^, *b)*.

Methods. Formulae to compute the standard error of the difference and standard deviation of the difference between a subtest and a client's mean subtest scores were applied to determine critical values for reliable and abnormal differences. The data used were derived from the TEA standardization sample (N = 154).

Results. Tables for examining whether an individual's TEA subtest profile contains reliable and abnormal subtest discrepancies are presented.

Conclusions. Elegant methods of analysing a subtest profile were extended for use with the Test of Everyday Attention. In keeping with the rationale underlying the measurement of neuropsychological deficit (Lezak, 1995), these methods complement the existing TEA *normative* comparison standards by providing *individual* comparison standards for a client's performance. Guidance on the use of the tables is offered; the distinction between reliable and abnormal differences is highlighted.

Attentional problems are a common and debilitating feature of many neurological and psychiatric disorders (e.g. Crawford, Parker & McKinlay, 1992; Lezak, 1995; Walsh, 1991). However, the assessment of attentional dysfunction is a problematic area in clinical neuropsychological practice as most existing tests do not reflect current theoretical knowledge on attention and its fractionation (e.g. Posner & Peterson, 1990) and in many cases are poorly standardized (Lezak, 1995).

* Requests for reprints.

The Test of Everyday Attention (TEA; Robertson, Ward, Ridgeway & Nimmo-Smith, 1994, 1996) was developed to improve upon existing methods of assessing attentional problems. It consists of eight subtests which measure sustained, selective and divided attention in visual and auditory modalities and was standardized on 154 healthy participants aged between 18 and 80. In addition to reflecting current thinking on the fractionation of attention it is designed to be ecologically valid, i.e.the subtests are designed to mimic everyday activities. Further details on the theoretical rationale behind the design of the subtests and provisional findings in clinical samples are provided in Robertson *et al.* (1996).

In neuropsychological assessment considerable emphasis is placed on analysing a client's pattern of relative strengths and weaknesses (Crawford, Parker & McKinlay, 1992; Lezak, 1995; Walsh, 1991). The TEA is well suited to this approach as the eight subtests were standardized on the same sample and, with the exception of the Elevator Counting subtest, are designed to yield equivalent scores (TEA subtests are scaled to be normally distributed with a common mean and standard deviation of 10 and 3 respectively).

The first aim of the present study is to provide a method of examining an individual's TEA profile for the presence *of reliable* differences between the subtests on the assumption that clinicians would wish to avoid interpreting differences which are likely simply to reflect measurement error. Any such attempt has to strike a compromise between control of the Type I and Type II error rates, i.e. the probability of incorrectly rejecting or accepting the null hypothesis which, in the present case, would be that a client's profile contains no subtest differences. For example, using the available data on the reliability of TEA subtests, one could readily calculate the difference between each TEA subtest and any other required for a specified level of significance. This is the approach adopted by Wechsler (1981) for the analogous purpose of examining the reliability of differences between subtests from the Wechsler Adult Intelligence Scale-Revised (WAIS-R).

It would be appropriate to use the above approach only if the clinician has made *an a priori* decision to examine the difference between *one* pair of subtests. However, in practice clinicians will want to compare more than one pair. Moreover, although one should seek to form hypotheses concerning a client's strength and weaknesses prior to formal testing, the client's history and behaviour at interview may provide insufficient grounds to generate such hypotheses. In this situation comparisons will bepotf *hoc* so that in effect all possible comparisons are made, even though the clinician will focus attention on the subtest comparisons which yielded the largest discrepancies. Therefore, the principal problem with this approach is that it does not introduce any correction for the inflation of the Type I error rate that occurs when multiple comparisons are conducted (Crawford, 1992; Silverstein, 1982).

An unsatisfactory solution to this problem would be to produce modified critical values using a Bonferroni correction to reflect the fact that 21 comparisons were involved.[1] This would be achieved by dividing the desired alpha (significance) level, e.g. .05, by 21 (the number of comparisons) and then by 2 (for two-tailed values). Thus the difference between any two subtests would have to be significant beyond

[1] *See across* for note.

p = .00119 to maintain the overall alpha level (termed the familywise error rate) at .05. It can be seen that, although such an approach controls for inflation of the Type I error rate, it would result in very low power to detect subtest differences, i.e. the Type II error rate would be unacceptably high (see Silverstein, 1982).

Silverstein (1982) and Knight & Godfrey (1984) independently suggested a method for use with the WAIS-R which achieves a useful compromise between the need to minimize both Type I and Type II errors. Instead of potentially comparing all subtests with each other, each subtest is compared with a client's mean subtest score. In the case of a full-length TEA this would reduce the number of comparisons to 7 so that, when a Bonferroni correction is applied to maintain the Type I error rate at the specified level, the loss of power to detect subtest differences would be much less acute, i.e. 7 would appear in the denominator rather than 21.

If a client exhibits *reliable* TEA subtest differences a second stage would be to determine whether such differences are also *abnormal,* i.e. are of a magnitude such that they would occur rarely in the healthy population. Silverstein *(I984a)* developed a formula to allow this further question to be addressed. As was the case for the formula for reliable differences, it is based on comparing each subtest with an individual's mean subtest score.

Method

Deriving critical values for reliable TEA subtest differences

To determine how large the difference between a particular TEA subtest and an individual's mean subtest score must be to achieve a given level of statistical significance requires calculation of the standard error of the difference (SEjiff). The SE^iff for each subtest was calculated using a formula developed by Davis (1959) and subsequently used by Silverstein (1982) and Knight & Godfrey (1984). Using Silverstein's notation the formula is:

$$\mathrm{SE}_{\mathrm{diff}} \sqrt{\left(\frac{k-2}{k}\right) S_1^2 + \frac{1}{k^2} \sum S_j^2}, \qquad (1)$$

where k = the number of subtests (seven in the present case when used with a full administration of the TEA), S^2i = the variance error of measurement for the subtest under consideration, and $\%S^2j$ = the summed variance errors of measurement for all subtests (including the subtest under consideration). The variance error of measurement for each subtest was obtained using the formula:

$$S_j^2 = \mathrm{SD}^2 \sqrt{1 - r_{jj}}, \qquad (2)$$

[1]The methods discussed assume that subtests have a common mean and standard deviation. The Elevator Counting subtest of the TEA does not meet this requirement as raw scores cannot be converted to scaled scores. Therefore, for present purposes a full length TEA will be treated as consisting of seven subtests. With seven subtests, a comparison of each subtest with all other available subtests yields a total of 21 comparisons.

where ryy = the reliability of the subtest as reported in the TEA manual (Robertson *et at.,* 1994, p. 6) and SD = 3, the standard deviation for TEA scaled scores. The reliabilities are alternate form reliability coefficients obtained from the subset of the TEA standardization sample (N = 118) who were administered version B following administration of the standard version (version A). The exception to this occurred in the case of the Lottery subtest where the alternate form reliability coefficient reported for a sample of 74 stroke patients was used; reliabilities were not calculated for this subtest in the standardization sample because of concern over ceiling effects (Robertson *et al.,* 1994).

To obtain critical values, the SE^iff for each subtest obtained from formula (1) was multiplied by the values of Z (the standard normal deviate) corresponding *to family wise* error rates of .15, .10, .05 and .01. For example, to determine the critical value for significance at the .05 level for the Lottery subtest, .05 was divided by 7 (the number of subtests and hence also the number of comparison to be made) and then by 2 (to obtain a two-tailed value); this yields a figure of 0.00357. The Z corresponding to this *p* value is 2.69. Multiplying the SE^ff for the Lottery subtest (1.35) by this latter figure yields a critical value of 3.63.

Deriving critical values for the abnormality of TEA subtest differences

Silverstein (19840) developed a formula to calculate the standard deviation of the difference between a mean subtest score and one of the subtests entering into that average. In order to use this method the TEA subtest correlation matrix was computed using the scaled subtest scores of the full TEA standardization sample. Silverstein's formula is as follows:

$$SD_{Da} = 3\sqrt{1+G\ -2T_a}\ , \tag{3}$$

where SD^ is the standard deviation of the difference for subtest *a*\ 3 is the (common) standard deviation of the subtests; G is the mean of all the elements in the standardization sample subtest correlation matrix (including unities in the diagonal) and T_a is the mean of the elements in a row or column *a* of the matrix (again including the diagonal). To estimate how large a difference would be obtained by 100 per cent of the general population, the standard deviation of the difference for each subtest $(SDj)_a$ is multiplied by the value of Z corresponding to a particular value of/? (Silverstein, 1984^; see Silverstein 1984^ for details on the derivation of the formula). In the present case the $SDj)_a$ for each subtest was multiplied by Z values of 1.44, 1.64, 1.96 and 2.58 in order to estimate the size of subtest discrepancy which must be exceeded to occur in less than 15, 10, 5, and 1 per cent of the healthy population respectively.

Computer program for use with short-form administrations of TEA

The tables of critical values developed here are designed to be used when all TEA subtests have been administered. There is no technical difficulty in extending the approach to allow it to be used when only a selection of TEA subtests have been given; see Crawford *etal.* (1997) and Crawford (1997) for examples of such extensions for use in the analogous situation where a short-form WAIS-R has been administered. However, even if attention is restricted to TEA short-forms consisting of between four and six subtests there are a prohibitive number (63) of unique subtest combinations. A computer program for PCs was therefore developed which implements the formulae presented above and calculates the reliability and abnormality of subtest deviations for a user-selected combination of TEA subtests.

Results

Table 1 presents critical values for assessing whether an individual exhibits *reliable* differences between their scaled score on a TEA subtest and their mean subtest scaled score. Critical values are presented for the .15, .10, .05 and .01 levels of significance; the standard errors of difference are also presented so that clinicians can potentially construct critical values for other significance levels if they wish. Table 2 can be used to determine the *abnormality* of the difference between a subtest scaled score and the mean subtest scaled score.

Table 1. Size of difference (regardless of sign) between each subtest and an individual's mean subtest score on the TEA required for statistical significance at the .15, .10, .05 and .01 levels; the standard error of the difference is also presented

		Critical value for a given *p*			
TEA subtest	SE_{diff}	.15	.10	.05	.01
Map Search	1.11	2.56	2.73	2.99	3.55
EC with Distraction*	1.48	3.41	3.64	3.99	4.73
Visual Elevator	1.48	3.41	3.64	3.99	4.73
EC with Reversal	1.59	3.65	3.89	4.27	5.07
Telephone Search (TS)	1.11	2.56	2.73	2.99	3.55
TS—Dual Task	1.72	3.97	4.22	4.64	5.50
Lottery	1.35	3.10	3.30	3.63	4.30

* EC = Elevator Counting.

Table 2. Size of difference (regardless of sign) between each subtest and an individual's mean subtest score on the TEA such that it would be expected to occur in less than 15, 10, 5 and 1 per cent of the healthy population; the standard deviation of the difference is also presented

		Percentage of healthy population			
TEA subtest	SD_{Da}	15	10	5	1
Map Search	2.61	3.76	4.29	5.12	6.74
EC with Distraction*	2.37	3.42	3.89	4.65	6.12
Visual Elevator	2.40	3.46	3.94	4.71	6.20
EC with Reversal	2.21	3.18	3.62	4.33	5.70
Telephone Search (TS)	2.18	3.14	3.57	4.27	5.62
TS–Dual Task	2.56	3.68	4.20	5.02	6.60
Lottery	2.46	3.54	4.03	4.81	6.34

* EC = Elevator Counting.

Discussion

The TEA manual provides guidance on comparing an individual's subtest scores with the norms for her/his age group. The present tables supplement this information on *normative* comparison standards by permitting the clinician to also employ *individual* comparison standards. This latter approach is in keeping with the rationale of deficit measurement outlined by Lezak (1991, 1995) and others (e.g. Crawford, 1992; McKinlay & Gray, 1992; Walsh, 1991). Lezak (1995) notes that, because of the large variability in premorbid cognitive abilities, a particular score can represent a serious degree of impairment for one individual whilst representing an entirely normal level of functioning for another. She therefore emphasizes the need to compare a client's performance on any test against their performance on the other measures administered. In conducting such a profile analysis it is important to have an empirical method to assist in the formation of judgments concerning the importance of any discrepancies. Clinicians have normally had limited opportunities to test individuals drawn from the general healthy population. Thus we may form seriously distorted impressions of what degree of subtest scatter in a profile constitutes normal limits (Crawford & Allan, 1996). Furthermore, as the TEA has only been published recently, clinicians will also have only limited experience with the test in clinical populations.

The present tables provide a rapid and straightforward means of quantitatively analysing a client's profile of strengths and weaknesses. It could be argued that the potential alternative approaches outlined in the introduction are not only unsatisfactory on psychometric/statistical grounds but also on practical grounds because of the amount of data they would generate. In reaching a formulation, clinicians have to integrate information from the TEA with a large body of quantitative and qualitative information from other sources (e.g. the history, behavioural observations, and other test scores, etc.) As others have suggested (e.g. Crawford & Allan, 1996; Kaufman, 1990; Silverstein, 1982), the approach adopted here for the TEA strikes a useful compromise between the need to keep information load at a reasonable level whilst retaining clinically significant attributes of the subtest profile.

Before illustrating the use of the present tables it should be stressed that when comparing a subtest with the average subtest score this latter measure includes the subtest under consideration. If this is not appreciated, i.e. if each subtest is compared with the mean of the other subtests, use of the tables will produce invalid results. It can also be seen that the procedure adopted is efficient as only one mean need be calculated for all comparisons.

Take the example of a client who obtained a mean score of 11.14 and scores of 6, 8 and 9 on the Telephone Search, Map Search and Elevator Counting with Distraction (ECWD) subtests respectively; scores on the remaining subtests will be ignored in the interests of brevity. The first step is to determine if any of these subtest scores are estimated to be *reliably* different from the client's mean. In the present example it can be seen from Table 1 that the difference between Telephone Search and the mean is reliably different at the .01 level; the difference, 5.14, exceeds the critical value of 3.55 for this level of significance. For the Map Search subtest, the difference of 3.14 exceeds the critical value of 2.99 for significance at the

.05 level but is smaller than the critical value (3.55) required for the .01 level. The Telephone Search and Map Search subtests therefore qualify as significant relative weaknesses for this client. In contrast, the ECWD difference (2.14) does not achieve significance even at the less stringent .15 level for which the critical value is 3.41 (Wechsler, 1981 and others have suggested that this level of significance is acceptable when working with individuals).

It is important to be clear on the meaning of a reliable difference obtained by the use of Table 1. If a subtest difference exceeds the critical value for significance at the .05 level then it is estimated that there is a greater than 95 per cent probability of it reflecting a genuine relative strength or weakness in the cognitive ability measured by the subtest; correspondingly there is a less than 5 per cent probability that the difference has solely arisen because of measurement error in the instruments. Determining if differences are reliable is the first step in profile analysis and have a number of practical applications (see Silverstein, 1982). For example, identification of reliable differences can provide guidance when designing tailored rehabilitation packages, whether the aim is to identify weaknesses which can be targeted for intervention or to identify strengths which can be taken advantage of when targeting or finding solutions to other difficulties. However, for most purposes, including this example, it would be advisable to proceed beyond identifying reliable differences to assess the additional issue of the degree of abnormality of such differences. This is particularly important where an assessment has been conducted for medico-legal purposes, e.g. in cases of personal injury litigation. The presence of a reliable difference should not be taken to imply the presence of acquired cognitive impairment as many healthy individuals will have reliable differences between their performance on these tasks.

Returning to the example profile, Table 2 reveals whether the discrepancy between the client's mean and the Telephone Search subtest is of a sufficient magnitude to be considered abnormal; less than 5 per cent of the general healthy population would be expected to exhibit this large a discrepancy, i.e. the client's discrepancy exceeds the value of 4.27. In the case of Map Search it can be seen that the subtest discrepancy would not be considered to be rare or abnormal. Thus, although this discrepancy was statistically significant (i.e. it is unlikely that it simply reflects measurement error), a discrepancy of this magnitude would not be uncommon in individuals free of acquired deficits, i.e. more than 15 per cent of the healthy population would be expected to exhibit discrepancies on this subtest which are larger than that of the client. It would not be appropriate to examine the abnormality of the ECWD discrepancy as it did not qualify as a reliable difference.

Table 2 can be used to either generate clinical hypotheses concerning the presence of acquired deficits or to test previously derived hypotheses. In the latter situation much more in the way of convergent evidence from other sources would be required to infer the presence of deficits in cases where the level of abnormality is moderate (i.e. approaches or just exceeds the 15 per cent level) than in cases where the discrepancy is very extreme.

With a new test it can be difficult to form clear hypotheses regarding a client's likely pattern of performance. However, as practitioners gain experience with the test, and

as more research information becomes available on typical patterns of strengths and weaknesses in particular clinical conditions, this process will become progressively easier. Where hypotheses have been generated, the relevant subtest deviations could be compared against *one-tailed* values rather than the two-tailed values presented here. In the case of the *reliability* of the subtest deviations, multiplying the standard errors of the difference (column 1 in Table 1) by Z values of 2.03, 2.19, 2.45 and 2.98 will yield critical values for the .15, .10, and .01 levels respectively; these Z values were derived in identical fashion to the two-tailed values except that the alphas were not divided by 2 (see Method section). The use of one-tailed values will increase statistical power but it should be noted that the logic of hypothesis testing dictates that a deviation in the *opposite* direction to that predicted must be ignored, regardless of its magnitude (e.g., see Howell, 1997).

To obtain 'one-tailed' values for the *abnormality* of subtest differences the standard deviations of the difference in Table 2 should be multiplied by Z values of 1.03, 1.28, 1.64 and 2.32. Take the example of a client who obtains a score on the Visual Elevator subtest which is 3.1 points *below* her/his mean subtest score; it would be expected that less than 10 per cent of the population would exceed a discrepancy of this magnitude in this direction (2.40 X 1.28 = 3.07, which is less than the obtained discrepancy). Tables of one-tailed values can be obtained from the first author.

To summarize the above points: Table 1 is concerned with the reliability of any subtest differences and, in line with Silverstein (1982), we suggest that it should be used as the first step in any profile analysis. If subtest differences are not reliable (i.e. they may simply be reflecting measurement error) then it would be unwise to draw any clinical inferences from them. Any subtest differences that are reliable can then be examined for their degree of abnormality if this additional issue is relevant to the assessment question(s). Consideration should be given to using one-tailed values where the clinician has formed hypotheses regarding strengths or weaknesses on particular subtests.

The tables presented here are designed for use with a full-length TEA. The validity of any inferences made from them will therefore be compromised if they are used with a short-form administration. In the case of the table for detecting reliable differences (Table 1), the reliability of the short-form mean against which a subtest is compared will differ from the reliability of the full-length mean; more importantly the Bonferroni correction used in construction of this table assumes that seven subtests were administered. In the case of the abnormality of differences (Table 2), the estimated rarity of a discrepancy from the mean score is a function of the averaged correlation between the subtest in question and the other subtests; thus omitting a subtest which has either a particularly high or low correlation with the subtest in question will markedly effect the size of the critical values.

There are, however, a number of factors which lead to the use of short forms in practice. For example, a client may have a motor or sensory disability which precludes the use of some TEA subtests. Similarly, client fatigue or simple time pressure may dictate the use of a short form. Finally, subtests may be selected by the clinician to test specific hypotheses generated from other test results, clinical observations or client self-reports. It was impractical to provide tabled values for all possible short

forms, nor did we consider it profitable to provide tables for selected short forms given the combination of factors which determine subtest selection in a clinical setting. For these reasons, and with the assumption that most if not all clinicians will have access to PCs, a program was developed which generates information equivalent to that contained in Tables 1 and 2 for any combination of TEA subtests. A compiled version of the program on a 3.25 disk is available free of charge from the first author.

References

Crawford, J. R. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford, D. M. Parker & W W McKinlay (Eds), *A Handbook of Neuropsychological Assessment,* pp. 21-49. Hove: Erlbaum.

Crawford, J. R. (1997). WAIS-R short forms: Assessing the statistical significance of subtest differences. *British Journal of Clinical Psychology,* 36, 601-608.

Crawford, J. R. & Allan, K. M. (1996). WAIS-R subtest scatter: Base rate data from a healthy UK sample. *British Journal of Clinical Psychology,* 35, 235-247.

Crawford, J. R., Allan, K. M., McGeorge, E & Kelly, S. M. (1997). Base rate data on the abnormality of subtest scatter for WAIS-R short-forms. *British Journal of Clinical Psychology,* 36, 443-444.

Crawford, J. R., Parker, D. M. & McKinlay, W W (Eds) (1992). *A Handbook of Neuropsychological Assessment.* Hove:Erlbaum.

Davis, F. B. (1959). Interpretation of differences among average and individual test scores. *Journal of Educational Psychology,* **50,** 162-170.

Howell, D. C. (1997). *StatisticalMethods for Psychology*, 4th ed. Boston, MA: PWS-Kent.

Kaufman, A. S. (1990). *Assessing Adolescent and Adult Intelligence.* Boston MA: Allyn & Bacon.

Knight, R. G. & Godfrey, H. P D. (1984). Assessing the significance of differences between subtests on the Wechsler Adult Intelligence Scale-Revised. *Journal of Clinical Psychology,* **40,** 808-810.

Lezak, M. D. (1991). Identifying neuropsychological deficits. In R. G. Lister & H. J. Weingartner (Eds), *Perspectives on Cognitive Neurosaence,* pp. 357-367. Newark: Oxford University Press.

Lezak, M. D. (1995). *Neuropsychological Assessment,* 3rd ed. New^brk: Oxford University Press.

McKinlay, W W & Gray, J. M. (1992). Assessment of the severely head-injured. In J. R. Crawford, D. M. Parker & W W McKinlay (Eds), *A Handbook of Neuropsychological Assessment,* pp. 363-378. Hove: Erlbaum.

Posner, M. I. & Peterson, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience,* 13, 25-42.

Robertson, I. H., Ward, T., Ridgeway, V & Nimmo-Smith, I. (1994). *The Test of Every day Attention.* Bury St Edmunds: Thames Valley Test Company.

Robertson, I. H., Ward T, Ridgeway, V & Nimmo-Smith, I. (1996). The structure of normal human attention: The Test of Everyday Attention. *Journal of the International Neuropsychological Society, 2,* 525-534.

Silverstein, A. B. (1982). Pattern analysis as simultaneous statistical inference. *Journal of Consulting and Clinical Psychology,* **50,** 234-240.

Silverstein, A. B. (1984^). New formulas for evaluating the abnormality of test score differences. *Journal of Psychoeducational Assessment,* 2, 79-82.

Silverstein, A. B. (1984£). Pattern analysis: The question of abnormality. *Journal of Consulting and Clinical Psychology,* **52,** 936-939.

Walsh, K. W (1991). *Understanding Brain Damage,* 2nd ed. Melbourne: Churchill Livingstone.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised.* New^brk: Psychological Corporation.

Received 12 November 1996; revised version received 21 March 1997